Geometric Comparison of Popular Mixture-Model Distances

Scott A. Mitchell [*]

Abstract

Statistical Latent Dirichlet Analysis produces mixture model data that are geometrically equivalent to points lying on a regular simplex in moderate to high dimensions. Numerous other statistical models and techniques also produce data in this geometric category, even though the meaning of the axes and coordinate values differs significantly. A distance function is used to further analyze these points, for example to cluster them. Several different distance functions are popular amongst statisticians; which distance function is chosen is usually driven by the historical preference of the application domain, information-theoretic considerations, or by the desirability of the clustering results.

Relatively little consideration is usually given to how distance functions geometrically transform data, or the distances algebraic properties. Here we take a look at these issues, in the hope of providing complementary insight and inspiring further geometric thought. Several popular distances, χ^2, Jensen - Shannon divergence, and the square of the Hellinger distance, are shown to be nearly equivalent; in terms of functional forms after transformations, factorizations, and series expansions; and in terms of the shape and proximity of constant-value contours. This is somewhat surprising given that their original functional forms look quite different. Cosine similarity is the square of the Euclidean distance, and a similar geometric relationship is shown with Hellinger and another cosine. We suggest a geodesic variation of Hellinger. The square-root projection that arises in Hellinger distance is briefly compared to standard normalization for Euclidean distance. We include detailed derivations of some ratio and difference bounds for illustrative purposes. We provide some constructions that nearly achieve the worst-case ratios, relevant for contours.

Please see the enclosed technical report for more details; here are some highlights.

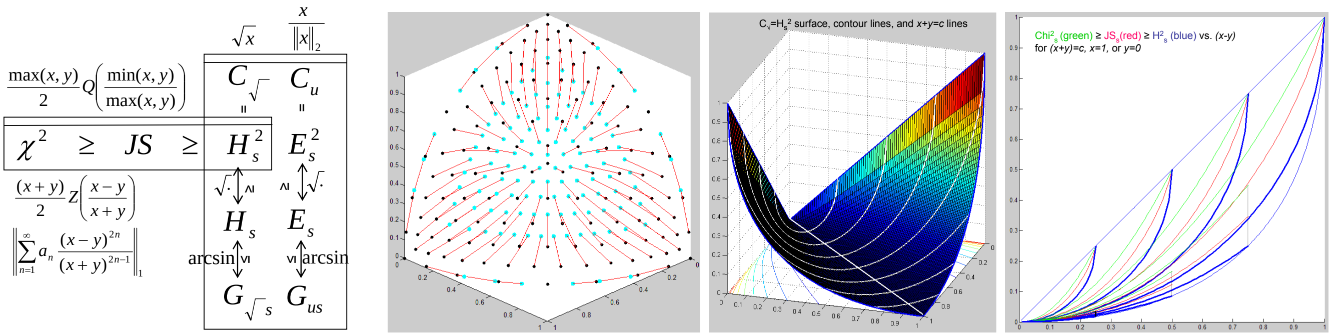

Figure 1: Left, distance metric taxonomy. Left-center, the relationship between the images of uniform points on the 3-simplex under standard normalization (small-black) and square-root (large-cyan). Lines connect the two images of the same point. Note the fixed-point at each sub-simplex center. Right-center, one-dimensional H^2, Right, comparison of the one-dimensional χ^2, JS and H_s^2.

Figure 1 summarizes the results of the paper. Here C is cosine similarity. χ^2 is Chi-squared. JS is Jensen-Shannon (a.k.a. half the Jeffreys Divergence). H is Hellinger. E is Euclidean. G is geodesic distance on the unit sphere. The double-headed arrows denote a global order-preserving isomorphism. The χ^2, JS, and H_s^2 inequalities hold componentwise, before taking norms. The equations above and below denote their common functional forms after transformations of variables.

Specifically, we show that each of the χ^2, JS, and H^2 distances can be factored into

$$d(x,y) = \|\frac{p}{2}Q(q)\|_1 = \|\frac{u}{2}Z(z)\|_1 = 1 - \|\frac{u}{2}W(z)\|_1.$$

[*] samitch@sandia.gov, Computer Science and Informatics, Sandia National Laboratories

Here $p = x + y$, $d = x - y$ and $q = d/p$; also $u = \max(x, y)$, $v = \min(x, y)$, and $z = v/u$. All the Q and Z are similar: $Q(q) = \sum_{n=1}^{\infty} a_n q^{2n}, 1 \geq a_n > 0$; all are monotonic. Moreover we prove bounded ratio and bounded difference. The Q functions are increasing, and the ratio of Q functions is increasing. This is equivalent to the Z functions and their ratios decreasing. We provide constructions that nearly achieve the extremes.

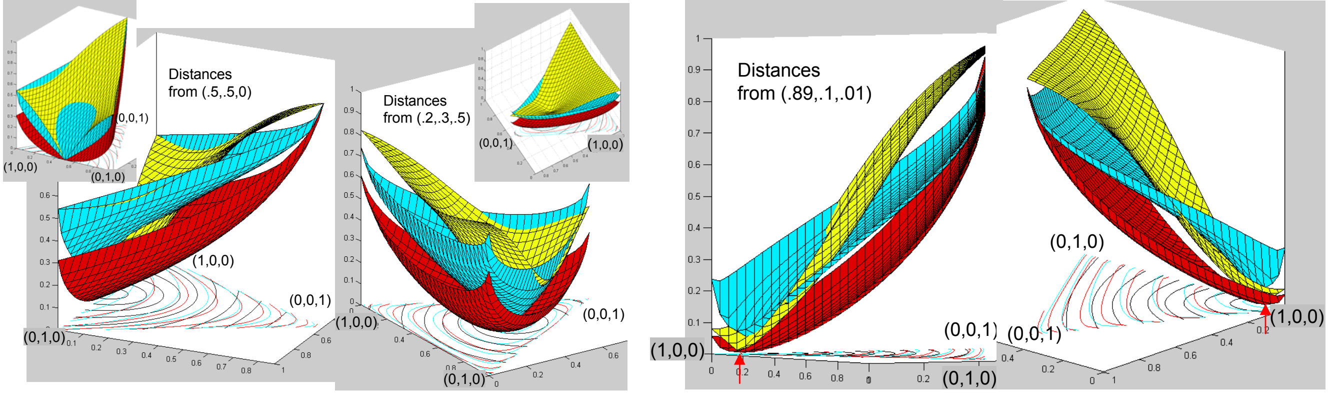

Figure 2: Euclidean(yellow), Hellinger(blue), and JS(red) distances on 3d mixture models. Note the shape of contour lines. H_s and JS_s are steeply sloped near $(1, 0, 0), (0, 1, 0)$.

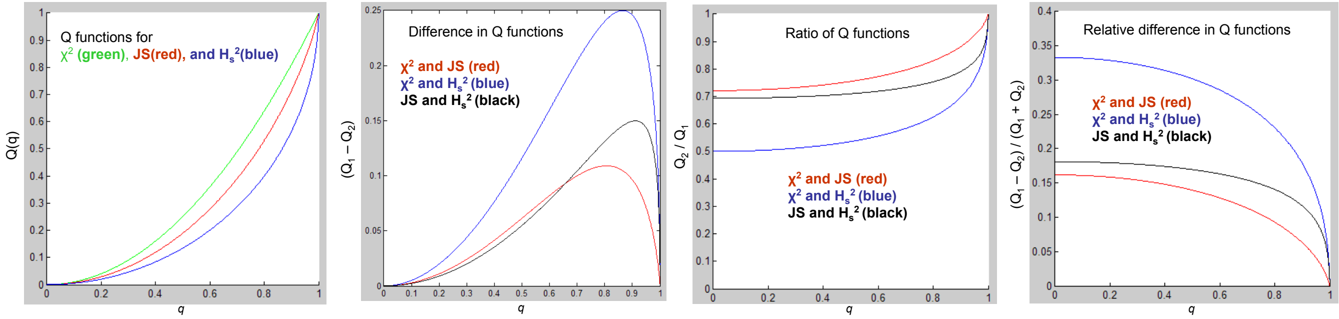

Figure 3: Graphs of relationships between the Q functions.

k	ϵ	a	b	χ^2	$\frac{JS}{\chi^2}(x,y)$	$\frac{JS}{\chi^2}(x,z)$	$\frac{H_s^2}{\chi^2}(x,y)$	$\frac{H_s^2}{\chi^2}(x,z)$	$JS(x,z')$	$\frac{H_s^2}{JS}(x,z')$	$\frac{H_s^2}{JS}(x,z')$
∞	$\to 0$	$\to 0$	$\to 0$	$\to 0$	.721	1	.5	1	$\to 0$	.693	1
5	.01	.26	.24	.00160	.7215	.998	.5002	.997	.00115	.6932	.9989
5	.08	.33	.17	.102	.73	.91	.51	.83	.075	.70	.92
5	.16	.41	.09	.41	.78	.95	.57	.91	.320	.72	.97
9	.08	.205	.045	.41	.78	.998	.57	.996	.320	.72	.957

Figure 4: Near worst-case ratio constructions for contours.

	Q^*	q^*	Z^*	z^*			Q^*	q^*	Z^*	z^*	R bound
$R_{\chi-H}$	$1/2 =$ .500	0	$1/2 =$ .500	1		$D_{\chi-H}$	.250	.866	.270	.087	.5
$R_{\chi-JS}$	$1/2 \log 2 >$ .721	0	$1/2 \log 2 >$ .721	1		$D_{\chi-JS}$	.110	.807	.122	.127	.279
R_{JS-H}	$\log 2 >$ .693	0	$\log 2 >$ .693	1		D_{JS-H}	.150	.912	.158	.055	.307
	max is 1 at $q = 1$ and $z = 0$					min is 0 at $q = 0$, $q = 1$, $z = 0$, and $z = 1$					

Figure 5: Left, exact minimum ratio between Q and Z functions, and the limit point where this is achieved. Right, provable maximum differences between Q and Z functions.